纵向题目作答时间模型：对潜在加工速度的发展追踪*

詹沛达** 陈琦鹏

（浙江师范大学教师教育学院，金华 321004）

摘　要　为实现对潜在加工速度发展的客观追踪，基于多元正态分布和潜在增长曲线提出了四个纵向题目作答时间(RT)模型。四个模型的测量模型一致，差异主要为描述潜在加工速度随时间变化的结构模型。实证研究结果表明四个模型均有实践可应用性，且它们的数据分析结果具有较高的一致性。模拟研究 1 表明四个模型在不同模拟条件下的参数估计返真性良好，且基于潜在增长曲线的纵向 RT 模型对潜在加工速度的估计精度略高于基于多元正态分布的纵向 RT 模型的。模拟研究 2 结果表明四个模型对中低比例(<60%)的随机缺失数据均具有一定的耐受性。总之，本文提出的四个纵向 RT 模型具有实践可应用性且心理计量学性能良好，不仅丰富了纵向 RT 数据的分析方法，还拓展了纵向模型的应用范围。

关键词　题目作答时间；纵向数据；对数正态作答时间模型；潜在增长曲线；项目反应理论

1. 引言

在心理与教育研究中，研究者通常对个体或群体在特定时间跨度中的认知或行为的发展变化感兴趣。这类研究的目标侧重于刻画每个个体的发展趋势和群体的平均变化轨迹。对潜在建构随时间发展的测量需要基于纵向研究设计对个体及其所在群体进行多次测量。相比于横断研究，纵向研究可以得到更有说服力的变量关系(e.g., 因果推论)论证(刘红云, 孟庆茂, 2003; 刘源等, 2022; 温忠麟, 2017)。目前，针对不同的观测变量类型和潜变量类型(连续或分类)研究者们提出了众多纵向数据分析模型，比如纵向 Rasch/IRT 模型(Andersen, 1985; Embretson, 1991; von Davier er al., 2011)、潜在增长曲线模型(Kaplan, 2000)和潜在转换分析模型(Collins & Lanza, 2010)等。近些年，随着测验情境复杂性的增加和对精准测量/追踪的追求，一些更复杂的纵向数据分析模型被提出，如增长混合模型(Muthen & Muthen, 2000)、纵向诊断分类模型(Zhan et al., 2019)、深度知识追踪模型(Piech et al., 2015)和分段潜变量增长模型(Kohli & Harring, 2013)等。尽管纵向模型本身并没有限制所分析的数据类型及所测量的潜在建构，但纵观已有研究可发现几乎所有纵向模型仅关注对传统题目作答结果(response accuracy, RA)数据(e.g., 答对答错或李克特式题目得分)的分析，忽略了其他模态数据，进而局限于追踪 RA 数据测量的心理建构(e.g., 潜在能力)的发展变化。

在智能时代背景下，随着计算机(网络)化测评的普及，除传统 RA 数据外，对诸如题目作答时间(response time, RT)等过程数据的采集已越发普遍(韩雨婷等, 2022; 刘耀辉等, 2022)。在心理与教育测评

* 国家自然科学青年基金(31900795)资助.

** 通讯作者：詹沛达, E-mail: pdzhan@gmail.com

中，RT 数据作为一种 RA 数据的补充或平行数据[1]，描述了个体解决单一问题的总耗时，可用于分析个体解决问题时的潜在加工速度。这在一定程度上打破了传统心理测量中对速度测验和难度测验的功能划分。另外，因 RT 数据“具有标准化数据结构，符合心理计量模型的建模与分析要求”(詹沛达, 2022, p2)，近些年受到了研究者们的广泛关注，开发了诸多 RT 模型(de Boeck & Jeon, 2019; 郭磊等, 2017; 詹沛达, 2018)。比如对数正态 RT 模型(lognormal RT model, LRTM) (van der Linden, 2006; Klein Entink, Fox et al., 2009)、多维 LRTM (詹沛达等, 2022)、Box-Cox 正态 RT 模型(Klein Entink, van der Linden et al., 2009)、变速 LRTM (Fox & Marianti, 2016)和一些关注速度-精度权衡的 RT 模型(e.g., Ferrando & Lorenzo-Seva, 2007)。但纵观已有研究可发现几乎所有 RT 模型都仅适用于分析横断测评数据，即仅能分析被试在单一时间点测验中的潜在加工速度，无法追踪个体潜在加工速度的发展轨迹。

《深化新时代教育评价改革总体方案》(中共中央, 国务院, 2020)明确指出应“改进结果评价，强化过程评价，探索增值评价，健全综合评价，充分利用信息技术，提高教育评价的科学性、专业性、客观性。”近些年，随着学测融合(assessment as learning)理念的普及，以学生为中心、以学习为中心的测评理念逐步得到认可，进而可提供及时反馈及干预的形成性学测项目逐渐受到人们的关注，如诊断性补救教学(王立君等, 2020; Tang & Zhan, 2021)、自适应学测系统(张华华, 汪文义, 2016; Zhang & Chang, 2016)和智能导学系统(Woolf, 2009)等。如图 1 所示，通常形成性学测项目会根据对个体在时间点 p ($p = 1, ..., P$)上 RA 数据的分析结果提供相应反馈和学习材料，然后在时间点 $p + 1$ 上对其再次测试，后再次提供反馈和学习材料，如此往复；最终，可以通过对多个时间点上 RA 数据(即纵向 RA 数据)的分析来刻画学生的发展轨迹(Chen et al., 2018; Wang, S., Yang et al., 2018; Zhan, 2020)。目前，随着计算机化测验的普及，一些形成性学测项目已经可以便捷地采集每个时间点上个体对每道题目的 RT 数据(即纵向 RT 数据) (e.g., Wang, S., Hu et al., 2020; Wang, S., Zhang et al., 2018)。Wang, S., Zhang 等人(2018)发现在自适应学测系统中，随着干预(反馈/学习)次数的增加，学生群体在下一个时间点上作答所有题目的平均 RT 会呈现下降趋势。Shi 等人(2018)发现在阅读理解任务中借助智能导学系统能够在一定程度上降低被试的 RT。而上述例子中导致观测变量 RT 降低的一个主要可能原因是被试的潜在加工速度随时间发生了增长。此时，如何合理分析纵向 RT 数据以实现对潜在加工速度发展的客观追踪，是一个兼具理论与实践意义的议题。

对此，Wang, S., Zhang et al. (2018)及 Wang, S., Zhang 和 Shen (2019)提出了动态 RT 模型。该模型假设个体潜在加工速度的变化是由当前时间点上个体是否掌握题目所需属性或其他协变量导致；而这一定程度上会限制该模型的实践应用。首先，该模型需要在认知诊断测评中与 RA 数据分析模型联合使用，但实践中单独关注 RT 数据的分析也很常见(e.g., Guo et al., 2021; van der Linden, 2006; Wang, C. et al., 2013; 詹沛达等, 2020)，且非认知诊断测评也可常采集 RT 数据。其次，该模型假设但也同时约束了潜

[1] RT 和题目作答结果包含有关同一个问题解决过程的平行信息，比如，某学生正确解决特定问题花费 10 秒。

在加工速度的变化原因，对一些测验时间点间隔比较长的纵向研究而言，这可能导致一些其他因素(如,自然成长和知识迁移)的影响被忽视。另外，实践中并不是所有研究都采集了协变量，也不是所有研究都对协变量的影响感兴趣。因此，仍有必要建构一些应用约束较少、适用场景更宽泛的纵向 RT 模型。

综上所述，已有的纵向数据分析模型主要聚焦对纵向 RA 数据的分析，缺乏对纵向 RT 数据的关注；且已有的 RT 模型多限于分析横断测评数据，无法追踪学生潜在加工速度随时间的发展。对此，本研究拟基于两类常见的纵向数据分析方法(i.e., 多元正态分布建模和潜在增长曲线建模)对最具代表性的 LRTM 进行拓展，提出四个纵向 RT 模型；以期实现对个体潜在加工速度发展的客观追踪并丰富纵向 RT 数据的分析方法。对此，下文将按如下逻辑撰写。首先，简单回顾横断 LRTM，并基于此提出四个纵向 RT 模型。其次，通过对一则有关空间旋转能力的纵向 RT 数据的分析，呈现新模型的实践表现。然后，通过两则模拟研究来探究新模型的心理计量学性能。最后，总结研究结果并讨论研究局限和展望。

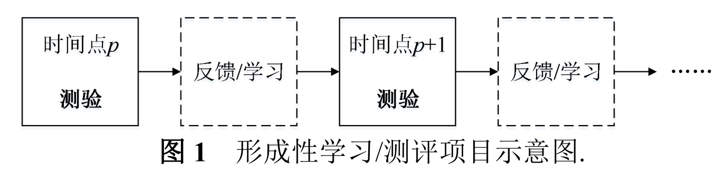


图 1　形成性学习/测评项目示意图.

2. 纵向题目作答时间模型

在心理计量模型中，纵向模型的一个核心作用是描述不同时间点上被试潜在建构的变化关系。根据描述方式的不同，通常可将纵向模型分类两类：一类是基于多元正态分布的纵向模型(e.g., Andersen, 1985; Embretson, 1991; Paek, Li, & Park, 2016; von Davier er al., 2011; Zhan et al., 2019)，另一类是基于潜在增长曲线的纵向模型(e.g., Bollen & Curran, 2006; Kaplan, 2000; Paek, Li, & Park, 2016; Wang, C., & Nydick, 2020)。前者类似于多维 IRT 模型，直接利用多元正态分布对被试在各时间点上的潜在建构进行建模，并可利用均值向量描述不同时间点上群体的发展轨迹；后者通过构建潜在建构与测验时间点之间的线性或非线性回归函数来描述潜在建构随时间点增加的变化趋势。

基于上述两种建模逻辑，本文拟提出两类纵向 RT 模型，分别为基于多元正态分布的纵向 RT 模型和潜在增长曲线纵向 RT 模型。进一步，基于不同的模型假设，本文在每类模型中再分别提出两个模型(即共四个模型)。从结构方程模型视角看，上述两类模型的差异在于描述各时间点上潜在构建关系的结构模型，而非测量模型。因此，下文先介绍统一的测量模型，然后再结合不同的结构模型逐一阐述四个新模型。

2.1. 模型建构

2.1.1. 测量模型

针对横断 RT 数据，LRTM 是目前最常用的 RT 测量模型之一。设定 T_{ni} 为被试 n (n = 1, ..., N)对题

目 i (i = 1, ..., I)的作答时间。则 LRTM 可表示为

$$\log T_{ni} = \xi_{\mathrm{i}} - \phi_i \tau_n + \varepsilon_{ni},\ \varepsilon_{ni} \sim N(0, \omega_i^{-2})\ , \tag{1}$$

或

$$\log T_{ni} \sim N(\xi_i - \phi_i \tau_n, \omega_i^{-2})\ , \tag{2}$$

其中，τ_n 是被试 n 的潜在加工速度；ξ_i 为题目时间强度参数，表示解答题目 i 所必需的时间；φ_i 为题目时间区分度参数，反映潜在加工速度对观察作答时间的影响程度；ε_{ni} 为残差，ω_i 是残差的标准差的倒数，可以将其视为题目时间峰度参数。

对于纵向测评而言，当整个测验包含 P 个测验时间点，则第 p 个时间点上纵向 LRTM 的测量模型可表示为：

$$\log T_{nip} \sim N(\xi_{ip} - \phi_{ip} \tau_{np}, \omega_{ip}^{-2})\ , \tag{3}$$

其中，T_{nip} 是时间点 p 上被试 n 对题目 i 的作答时间；ξ_{ip}、φ_{ip} 和ω_{ip} 分别是时间点 p 上题目 i 的时间强度参数、时间区分度参数和时间峰度参数；τ_{np} 是时间点 p 上被试 n 的潜在加工速度。

2.1.2. 基于多元正态分布的纵向题目作答时间模型

为描述 P 个时间点上τ_{np}之间的关系，一种最直接的方法是构建多元正态分布，如图 1(a)。即假设$\boldsymbol{\tau}_n = (\tau_{n1}, ..., \tau_{nP}\ )^{\mathrm{T}}$是遵循多元正态分布的多维潜在加工速度向量：

$$\boldsymbol{\tau}_n \sim MVN(\boldsymbol{\mu}, \Sigma) = MVN\left(\begin{pmatrix}\mu_1 \\ \vdots \\ \mu_P\end{pmatrix}, \begin{pmatrix}\sigma_{\tau 1}^2 & \cdots & \rho_{1P}\sigma_{\tau 1}\sigma_{\tau P} \\ \vdots & \ddots & \vdots \\ \rho_{P1}\sigma_{\tau P}\sigma_{\tau 1} & \cdots & \sigma_{\tau P}^2\end{pmatrix}\right), \tag{4}$$

式中，$\boldsymbol{\mu} = (\mu_1, ..., \mu_P\ )^{\mathrm{T}}$为 P 个时间点的潜在加工速度的均值向量；Σ为方差协方差矩阵，描述了 P 个时间点的潜在加工速度之间的关系。该模型直接估计的各个时间点上的潜在加工速度，因此可直接使用$\hat{\boldsymbol{\tau}}_n$描述被试个体潜在加工速度的发展轨迹。此时，可以用$\hat{\tau}_{n(p+1)} - \hat{\tau}_{np}$描述相邻时间点个体水平的变化程度，用$\hat{\mu}_{p+1} - \hat{\mu}_p$描述相邻时间点群体均值的变化程度。

实际上，该模型可视为多维 LRTM (詹沛达等, 2020)在纵向 RT 数据分析中的应用。因此，与多维 LRTM 一样，该模型中Σ的所有元素均需自由估计，即Σ中有 $P(P + 1)/2$ 个待估计参数。该做法相对优点是考虑了所有时间点上潜在加工速度之间的相互影响，相对缺点是当时间点 P 数量较多时参数估计计算量较大且易出现估计不收敛问题。

为缩减待估计参数数量，可通过引入马尔可夫性质(Markov property)来约束Σ中的待估计参数，如图 1(b)。目前已有许多研究将马尔可夫性质引入纵向数据分析中(e.g., de Haan-Rietdijk et al., 2017; Wang, S., Yang al., 2018; Zhan, 2020)。基于马尔可夫性质，可假设被试在时间点 p 的潜在加工速度只与其在时间点 $p - 1$ 的潜在加工速度有直接关系。对此，首先将Σ做如下转换：

$$\Sigma = \mathbf{S}\Omega\mathbf{S} \ , \tag{5}$$

$$\mathbf{S} = \begin{pmatrix} \sigma_{\tau 1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{\tau P} \end{pmatrix} , \tag{6}$$

$$\Omega = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1P} \\ \rho_{21} & 1 & \cdots & \rho_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{P1} & \rho_{P2} & \cdots & 1 \end{pmatrix} , \tag{7}$$

其中，$\mathbf{S}$ 为标准差矩阵，Ω为相关系数矩阵。然后，因只考虑相邻时间点之间的直接关系，所以只需将相关矩阵Ω中相邻时间点的相关系数$\rho_{(p-1)p}$作为待估参数；而跨时间点的相关系数不视为待估计参数，由各相邻时间点上的相关系数连乘而来：

$$\rho_{ab} = \rho_{a(a+1)}\rho_{(a+1)(a+2)}\cdots\rho_{(b-2)(b-1)}\rho_{(b-1)b} \ , \tag{8}$$

其中，ρ_{ab}为两个不相邻的两个时间点 a 和 b 之间的相关系数。比如，当相邻时间点之间的相关系数$\rho_{12} = 0.9$，$\rho_{23} = 0.7$，$\rho_{34} = 0.8$ 时，则有不相邻时间点之间的相关系数$\rho_{13} = \rho_{12}\rho_{23} = 0.9 \times 0.7 = 0.63$，$\rho_{14} = \rho_{12}\rho_{23}\rho_{34} = 0.9 \times 0.7 \times 0.8 = 0.504$，$\rho_{24} = \rho_{23}\rho_{34} = 0.7 \times 0.8 = 0.56$。此时，$\Sigma$中待估计参数数量由 $P(P+1)/2$ 缩减为 $2P-1$。

为便于阐述，下文将不包含马尔可夫性质的和包含马尔可夫性质的模型分别简称为 MVN-LRTM 和 MVN-LRTM-M。另外，在采用锚题设计和重复测量设计的情况下，可将第一时间点上所有被试的潜在加工速度的均值和方差分别约束为$\mu_1 = 0$ 和$\sigma^2_{\tau 1} = 1$ 以保证模型的可识别性(Paek, Li, & Park, 2016)。

2.1.3. 基于潜在增长曲线的纵向题目作答时间模型

为描述 P 个时间点上τ_{np}之间的关系，多元正态分布外的另一种方法是构建潜在增长曲线，如图 1(c)：

$$\tau_{np} = \pi_{0n} + \pi_{1n}(p-1) + \varepsilon_{np}, \ \varepsilon_{np} \sim N\left(0, \sigma^2_{\varepsilon p}\right) \ , \tag{9}$$[2]

$$\begin{pmatrix} \pi_{0n} \\ \pi_{1n} \end{pmatrix} \sim MVN\left(\begin{pmatrix} \mu_{\pi_0} \\ \mu_{\pi_1} \end{pmatrix}, \begin{pmatrix} \sigma^2_{\pi_0} & \rho_{\pi_0\pi_1}\sigma_{\pi_0}\sigma_{\pi_1} \\ \rho_{\pi_1\pi_0}\sigma_{\pi_1}\sigma_{\pi_0} & \sigma^2_{\pi_1} \end{pmatrix}\right) \ , \tag{10}$$

式中，π_{0n}为被试 n 的截距系数，表示被试 n 的初始潜在加工速度水平；π_{1n}为被试 n 的增长系数，表示被试 n 的潜在加工速度随时间变化的程度；π_{0n}和π_{1n}服从二元正态分布，两者的均值μ_{π_0}和μ_{π_1}分别代表群体潜在加工速度的均值和群体潜在加工速度的平均增长率，方差协方差矩阵则描述了潜在加工速度的初始水平和增长系数之间的关系：$\rho_{\pi_1\pi_0} > 0$ 意味着初始水平越高的被试，其潜在加工速度随时间的增幅越大，反之则反；ε_{np}为残差。与 MVN-LRTM 不同，该模型没有直接估计各时间点上的τ_{np}，而是估计了每个被试的增长曲线系数(i.e., π_{0n}和π_{1n})；此时，可以用$\hat{\pi}_{1n}$描述相邻时间点个体水平的变化

---

[2] 也有研究不考虑残差项(e.g., Curtis, 2010)，即$\tau_{np} = \pi_{0n} + \pi_{1n}(p-1)$；预研究结果表明不考虑残差项的模型对实证数据的拟合结果较差.

程度，用$\hat{\mu}_{\pi_1}$描述相邻时间点群体均值的变化程度。

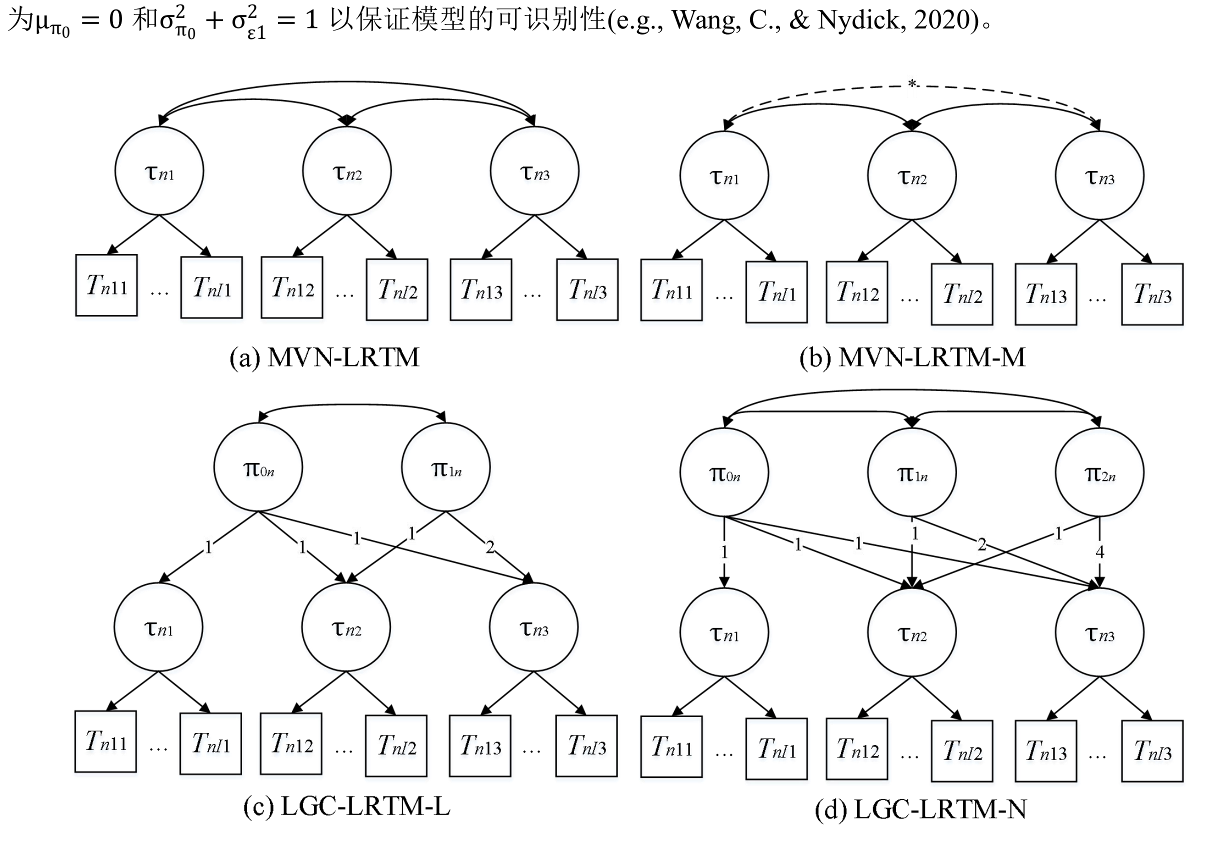
公式 9 假设τ_{np}随测验时间点的增加呈线性增长，而现实中τ_{np}随测验时间点的增加也可能呈非线性增长。此时，可在公式 9 中增加二次增长项来实现对潜在加工速度的非线性变化的描述，如图 1(d)：

$$\tau_{np}=\pi_{0n}+\pi_{1n}(p-1)+\pi_{2n}(p-1)^2+\varepsilon_{np},\ \varepsilon_{np}\sim N\left(0,\sigma^2_{\varepsilon p}\right)\ , \tag{11}$$

$$\begin{pmatrix}\pi_{0n}\\ \pi_{1n}\\ \pi_{2n}\end{pmatrix}\sim MVN\left(\begin{pmatrix}\mu_{\pi_0}\\ \mu_{\pi_1}\\ \mu_{\pi_2}\end{pmatrix},\begin{pmatrix}\sigma^2_{\pi_0} & \rho_{\pi_0\pi_1}\sigma_{\pi_0}\sigma_{\pi_1} & \rho_{\pi_0\pi_2}\sigma_{\pi_0}\sigma_{\pi_2}\\ \rho_{\pi_1\pi_0}\sigma_{\pi_1}\sigma_{\pi_0} & \sigma^2_{\pi_1} & \rho_{\pi_1\pi_2}\sigma_{\pi_1}\sigma_{\pi_2}\\ \rho_{\pi_2\pi_0}\sigma_{\pi_2}\sigma_{\pi_0} & \rho_{\pi_2\pi_1}\sigma_{\pi_2}\sigma_{\pi_1} & \sigma^2_{\pi_2}\end{pmatrix}\right)\ , \tag{12}$$

式中，π_{2n}为被试 n 的二次增长系数，其余参数同上。

实际上，这两个模型可视为变速 LRTM (Fox & Marianti, 2016)在纵向 RT 数据分析中的应用。当然，除包含二次增长项外，非线性增长模型中还可以进一步包含三次增长项或自由估计时间参数(Meredith & Tisak, 1990; Paek, Li, & Park, 2016)，但限于篇幅原因本文暂不关注它们。为便于阐述，下文将基于线性增长曲线和基于非线性增长曲线的模型分别称为 LGC-LRTM-L 和 LGC-LRTM-N。另外，在采用锚题设计和重复测量设计的情况下，可将第一时间点上所有被试的潜在加工速度的均值和方差分别约束为$\mu_{\pi_0}=0$ 和$\sigma^2_{\pi_0}+\sigma^2_{\varepsilon 1}=1$ 以保证模型的可识别性(e.g., Wang, C., & Nydick, 2020)。



(a) MVN-LRTM (b) MVN-LRTM-M
(c) LGC-LRTM-L (d) LGC-LRTM-N

图 1 四个纵向题目作答时间模型示意图($P = 3$).

注：虚线表示非相邻时间点之间的相关；“*”表示该相关系数由相邻时间点相关系数连乘得到.

2.1.4. 四个纵向题目作答时间模型的对比

表 1 呈现了本文所提出的四个纵向题目作答时间模型的对比(以 $P = 3$ 为例)。其中，当 $p = 1$ 时，有 $\tau_{n1} = \pi_{0n}$ 和 $\mu_1 = \mu_{\pi 0}$，即在起始点所有模型是完全等价的。其次，如上文所述，基于多元正态分布的两个模型和基于潜在增长曲线的两类模型在追踪学生发展时侧重点不同。具体而言，前者直接估计被试在各时间点上潜在加工速度的水平，并未直接关注潜在加工速度随时间的变化过程；而后者则估计被试潜在加工速度随时间的(线性或非线性)增长曲线系数，没有直接估计被试在各时间点上潜在加工速度的水平(可以计算出)。再次，对于纵向研究中可能出现的马太效应(e.g., von Davior et al., 2011; Zhan et al., 2019)，即被试之间的差异会随时间而增大，两类模型的描述视角也不一样。具体而言，前者直接估计群体在各时间点上潜在加工速度的标准差，可根据 $\sigma_{\tau(p+1)}/\sigma_{\tau p}$ 是否大于 1 来判断是否存在马太效应，$\sigma_{\tau(p+1)}/\sigma_{\tau p} > 1$ 时存在马太效应，$\sigma_{\tau(p+1)}/\sigma_{\tau p} \approx 1$ 时则不存在；而后者可根据 $\rho_{\pi_1\pi_0}$ 是否大于 0 来判断，$\rho_{\pi_1\pi_0} > 0$ 则存在马太效应，$\rho_{\pi_1\pi_0} \approx 0$ 时则不存在。需要强调的是 MVN-LRTM-M 和两个 LGC-LRTM 适用于 $P \geq 3$ 的测验情境；而当 $P = 2$ 时，直接使用 MVN-LRTM 即可。

表 1. 四个纵向题目作答时间模型的对比($P = 3$).

模型	个体水平			群体水平		
	$p = 1$	$p = 2$	$p = 3$	$p = 1$	$p = 2$	$p = 3$
MVN-LRTM	τ_{n1}	τ_{n2}	τ_{n3}	μ_1	μ_2	μ_3
MVN-LRTM-M	τ_{n1}	τ_{n2}	τ_{n3}	μ_1	μ_2	μ_3
LGC-LRTM-L	π_{0n}	$\pi_{0n}+\pi_{1n}$	$\pi_{0n}+2\pi_{1n}$	$\mu_{\pi 0}$	$\mu_{\pi 0}+\mu_{\pi 1}$	$\mu_{\pi 0}+2\mu_{\pi 1}$
LGC-LRTM-N	π_{0n}	$\pi_{0n}+\pi_{1n}+\pi_{2n}$	$\pi_{0n}+2\pi_{1n}+4\pi_{2n}$	$\mu_{\pi 0}$	$\mu_{\pi 0}+\mu_{\pi 1}+\mu_{\pi 2}$	$\mu_{\pi 0}+2\mu_{\pi 1}+4\mu_{\pi 2}$

2.2. 参数估计

本研究使用全贝叶斯马尔可夫链蒙特卡洛(MCMC)算法对四个纵向 RT 模型进行参数估计，并基于 JAGS (Ver 4.3.0) (Plummer, 2015)实现。相应的 JAGS 示例代码见 http://....。根据已有数据分析经验以及已有研究结果(詹沛达等, 2020; Fox & Marianti, 2016; Wang, S., Zhang et al., 2018), 本文选取了特定的先验分布。网络版附录 S1 章节中呈现了模型参数估计对高、中和低信息先验分布的稳健性分析结果, 结果表明四个新模型对包含不同信息量的先验分布均具有较高的稳健性。关于如何使用 JAGS 进行贝叶斯 MCMC 参数估计，可参阅 Curtis (2010)及 Zhan, Jiao, Man 和 Wang (2019)。

3. 实证数据分析

3.1. 数据描述与分析

本研究以一则有关空间旋转能力的自适应学测数据(Wang, S., Yang et al., 2018)为例来展现所提出模型的实践可应用性。该数据集包含 350 名被试在 5 个时间点上对 50 道题目(即每个时间点 10 道题目)的作答数据。具体而言，为平衡题目位置效应，该测验采用拉丁方设计(见表 2)，测验总共包含 5 个组块，每个组块包含 10 题(共 50 题)，并根据施测顺序形成 5 个版本的测验。在每个时间点施测时，每位学生随机抽取其中 1 个版本的测验。该数据已经被一些研究用于探究学生的学习轨迹(Chen et al., 2018; Wang, S., Yang et al., 2018; Wang, S., Zhang et al., 2019)。本研究拟分析该数据集中的 RT 数据来追踪被试潜在加工速度的发展。

实际上，该测验本质是一个采用了重复测量设计的纵向测验，只不过每名被试由于施测设计导致其在每个时间点上只作答了 10 道题目(1 个组块)，缺失另外 40 道题目(i.e., 设计缺失[missing by design])。因此，可将该数据重新整理为 350 人在 5 个时间点上共 250 道题目(每个时间点 50 题)上的纵向数据；其中，由设计缺失导致的缺失数据被视为完全随机缺失。图 2 呈现了 50 道题目的对数 RT 随时间变化趋势(剔除缺失值)，可发现明显的下降趋势。

分别使用 MVN-LRTM、MVN-LRTM-M、LGC-LRTM-L 和 LGC-LRTM-N 作为数据分析模型。四个模型均使用两条马尔可夫链，均预热 10,000 次，采样 5,000 次[3]。使用潜在量尺缩减因子(PSRF; Brooks & Gelman, 1998) 对作为 MCMC 算法的收敛指标，通常 PSRF < 1.1 或 1.2 表示参数估计已收敛。使用后验预测模型检验(posterior predictive model checking, PPMC)来评估模型−数据绝对拟合，其中后验预测概率(posterior predictive probability, *ppp*)接近 0.5 表明模型与数据拟合。本研究使用测验统计量(test statistics) (即仅关注真实数据与预测数据之间的差异，不涉及具体模型参数) (Levy & Mislevy, 2016)作为 PPMC 的差异测度。使用-2LL (-2×log likelihood)和 DIC (deviance information criterion) (Spiegelhalter et al., 2002)作为模型-数据相对拟合指标，指标值越小说明模型和数据拟合的越好。前者不包含模型复杂性惩罚，单纯反映模型与数据的拟合情况；而后者包含模型复杂性惩罚，在反映模型与数据拟合情况的同时还考虑了实践应用中的简约原则(parsimony principle) (Beck, 1943).

表 2. 实证研究的拉丁方设计.

测验版本	测验顺序($P = 5$)				
	$p = 1$	$p = 2$	$p = 3$	$p = 4$	$p = 5$
版本 1	组块 1	组块 2	组块 3	组块 4	组块 5
版本 2	组块 2	组块 3	组块 4	组块 5	组块 1
版本 3	组块 3	组块 4	组块 5	组块 1	组块 2

[3] 该设定下 MVN-LRTM 中潜在加工速度的方差协方差矩阵中部分元素未达到收敛标准(PSRF < 1.2)，随后将每条链迭代次数增加至 100,000 次(预热 90,000)，这些参数仍未完全达到收敛标准；其余参数均达到收敛标准。

版本 4	组块 4	组块 5	组块 1	组块 2	组块 3
版本 5	组块 5	组块 1	组块 2	组块 3	组块 4

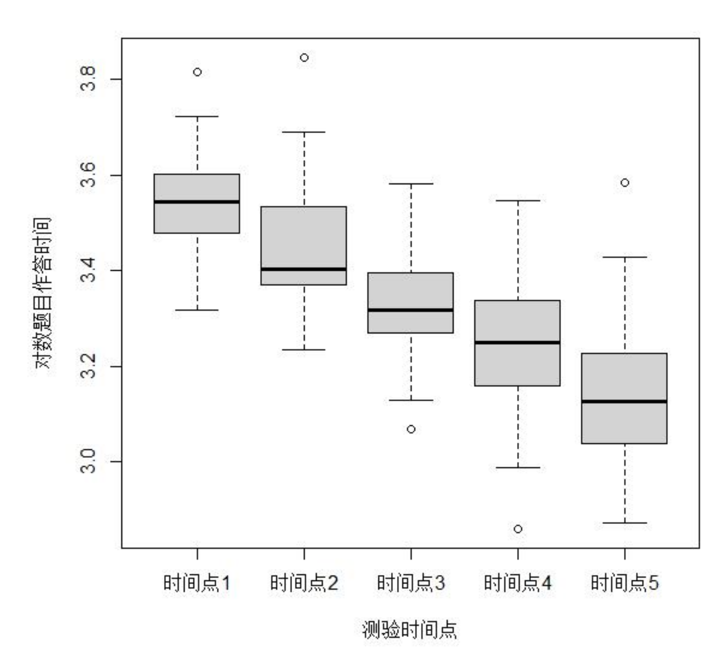


图 2 实证研究 5 个时间点上 50 道题目的对数题目作答时间分布.

3.2. 结果

需要强调的是，由于 MVN-LRTM 中潜在加工速度的方差协方差矩阵中部分元素没有达到收敛标准(PSRF < 1.2)，所以该模型与数据的拟合结果仅供参考(其余参数均达到收敛标准)；可能的原因是五元正态分布较难实现稳健的参数估计(e.g., Cai, 2010)，且该数据中样本量较小还包含较大比例的缺失值。其他三个模型的所有模型参数均达到收敛标准。

表 3 呈现了四个模型对实证数据的拟合情况。首先，根据各时间点上的 *ppp* 值，表明四个模型均拟合该数据。其次，不考虑 MVN-LRTM 时，剩余三个模型的对数据的相对拟合比较接近。其中，-2LL 指标值表明，在不考虑模型复杂性惩罚的前提下，LGC-LRTM-N 对该数据的拟合相对最好，即该模型得到的参数估计值相对最能反映数据的特征。而 DIC 指标值表明 MVN-LRTM-M 对该数据拟合相对最好，LGC-LRTM-L 次之且和 LGC-LRTM-N 几乎没有差异。总之，在简约原则下推荐使用 MVN-LRTM-M 分析该数据；但单纯从反映数据本身特征的角度看，LGC-LRTM-N 的拟合最好。

图 3 呈现了四个模型中所有被试潜在加工速度随时间的变化趋势(含群体均值变化)。首先，对任何模型而言，潜在加工速度的群体均值都呈较明显的增长趋势。具体而言，MVN-LRTM 的潜在加工速度均值向量$\boldsymbol{\mu} = (0, 0.297, 0.728, 0.996, 1.384)^T$，MVN-LRTM-M 的潜在加工速度均值向量$\boldsymbol{\mu} = (0, 0.311,$

0.757, 1.030, 1.393)^T，LGC-LRTM-L 的潜在加工速度均值向量$\boldsymbol{\mu}$ = (0, 0.433, 0.866, 1.299, 1.732)^T，LGC-LRTM-N 的潜在加工速度均值向量$\boldsymbol{\mu}$ = (0, 0.483, 0.955, 1.416, 1.867)^T。其次，同一类增长模型的变化趋势更接近，不同类模型之间的略有差异：两个 LGM-LRTM 对各时间点上群体均值的估计值大于两个 MVN-LRTM 对各时间点上群体均值的估计值。总之，被试潜在加工速度随时间增长的趋势可以较好地解释图 2 中 RT 随时间的下降趋势。

图 4 呈现了四个模型中所有时间点上潜在加工速度的估计值之间的相关系数图。可以看到，无论是同一模型对 5 个时间点上潜在加工速度的估计值之间，还是不同模型对同一时间点上潜在加工速度的估计值之间，均呈现高程度相关。一方面表明不同模型的估计值之间具有高度一致性，另一方面表明不同时间点上潜在加工速度之间也具有高度一致性(主要原因是该测验中各时间点之间的间隔较短)。

图 5 呈现了四个模型的题目参数估计值。因为该测验采用重复测量设计，所以仅有 50 道题目的题目参数。首先，四个模型的题目参数估计值之间具有较高的一致性，尤其是时间强度参数和时间峰度参数。其次，同一类模型的时间区分度参数估计值相对更接近。由于 LRTM 中时间区分度参数与潜在加工速度之间存在交互，导致两类模型的时间区分度参数估计值之间存在细微差异的可能原因是两类模型对潜在加工速度估计值存在细微差异性。具体而言，因为 LGC-LRTM 对潜在加工速度的估计值略大于 MVN-LRTM 的，所以 LGC-LRTM 对时间区分度参数的估计值略小于 MVN-LRTM 的。

此外，MVN-LRTM 和 MVN-LRTM-M 可以计算潜在加工速度随时间进展的量尺变化(i.e., $\sigma_{\tau(p+1)}/\sigma_{\tau p}$)，见表 4。可发现，在该测验中被试不存在马太效应，且被试之间的差异随时间进展还略微减小。另外，LGC-LRTM-L 中$\hat{\rho}_{\pi_1\pi_0} = -0.051$ (i.e., 被试增长系数与初始值成极弱负相关)也印证了该结论。

综上所述，实证研究结果表明四个纵向 RT 模型均具有实践可应用性且对同一批数据的分析结果具有较高的一致性。当然，由于实证数据分析主要用于呈现新模型的实践可应用性，其他一些数据本身相关的结论(e.g., 导致发展的原因)不再探讨。

表 3. 实证研究中模型-数据拟合结果.

模型	-2LL	DIC	*ppp_1*	*ppp_2*	*ppp_3*	*ppp_4*	*ppp_5*
MVN-LRTM	39123.936	39703.615	0.381	0.445	0.526	0.593	0.592
MVN-LRTM-M	39094.366	39872.355	0.419	0.457	0.489	0.556	0.600
LGC-LRTM-L	39056.191	39965.569	0.463	0.378	0.632	0.496	0.576
LGC-LRTM-N	39051.008	39967.630	0.455	0.341	0.634	0.506	0.605

注: MVN-LRTM 的方差协方差矩阵中部分元素估计未收敛，结果仅供参考.

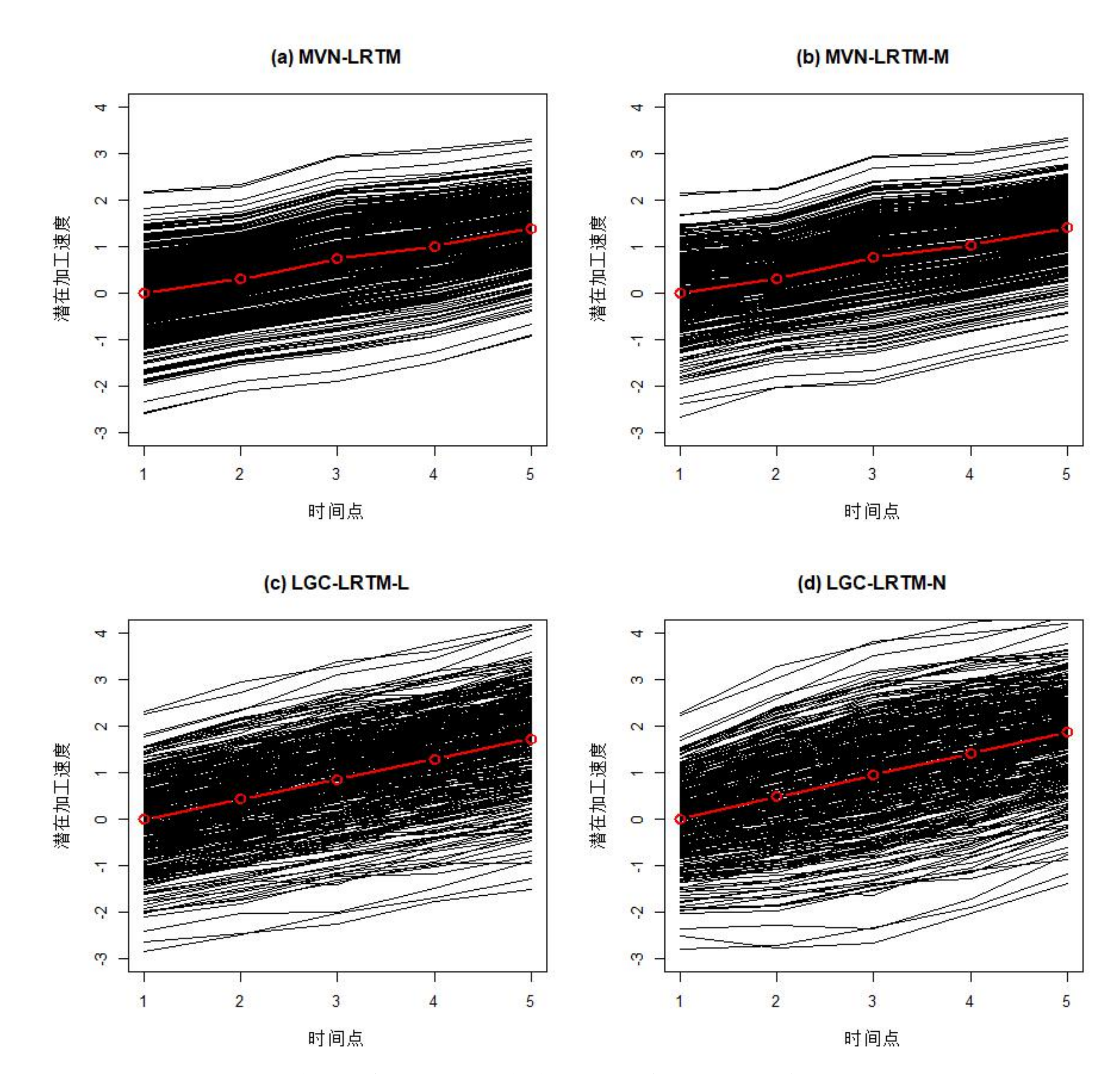


图 3. 实证研究中潜在加工速度随时间的变化趋势.

注: 红线为群体均值变化.

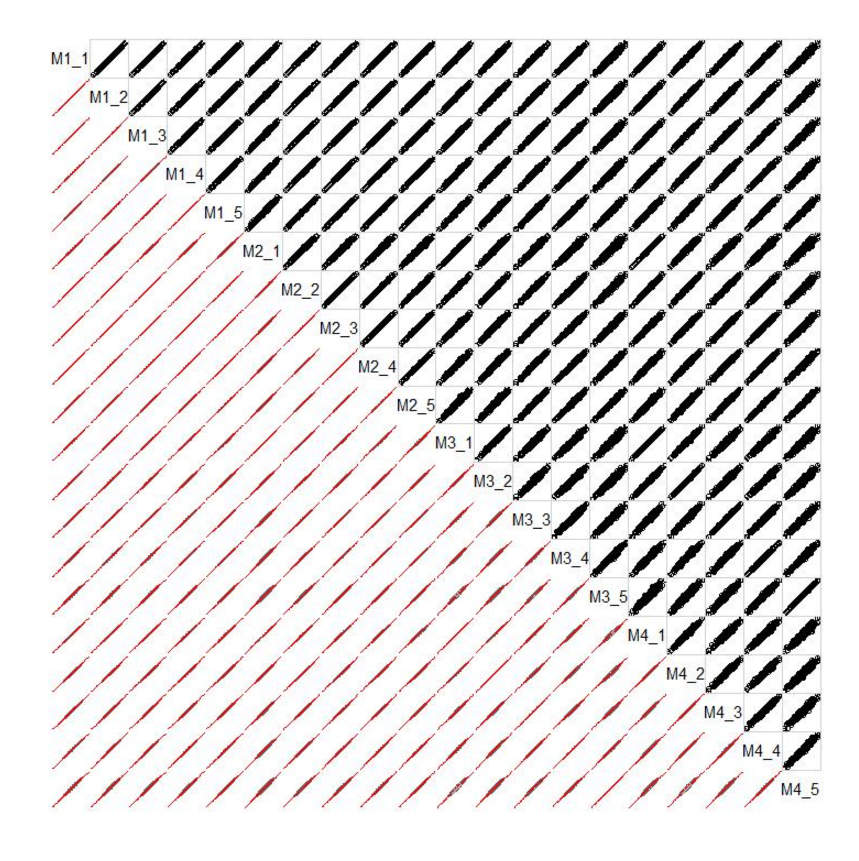


图 4. 实证研究中所有模型对所有时间点上潜在加工速度的估计值之间的相关系数图.
注: M1 = MVN-LRTM; M2 = MVN-LRTM-M; M3 = LGC-LRTM-L; M4 = LGC-LRTM-N; 下三角区域包含平滑拟合曲线和置信椭圆，上三角区域包含散点图.

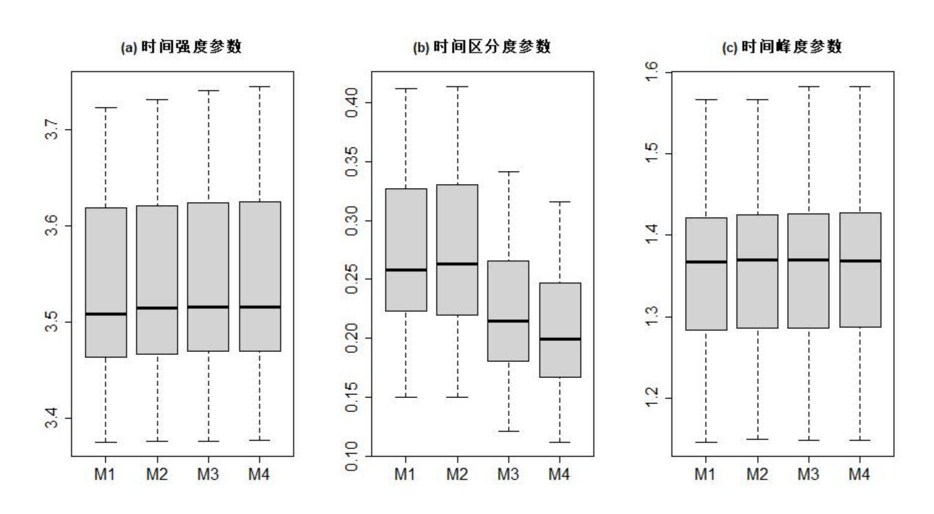


图 5. 实证研究中所有模型的题目参数估计值.
注: M1 = MVN-LRTM; M2 = MVN-LRTM-M; M3 = LGC-LRTM-L; M4 = LGC-LRTM-N.

表 4. 实证研究中潜在加工速度随时间进展的量尺变化.

模型	$\sigma_{\tau2}/\sigma_{\tau1}$	$\sigma_{\tau3}/\sigma_{\tau2}$	$\sigma_{\tau4}/\sigma_{\tau3}$	$\sigma_{\tau5}/\sigma_{\tau4}$
MVN-LRTM	0.933	1.096	0.948	0.925
MVN-LRTM-M	0.887	1.131	0.910	0.981

注: MVN-LRTM 的方差协方差矩阵中部分元素估计未收敛，结果仅供参考.

4. 模拟研究

实证数据分析表明新模型具有实践可应用性，下文通过两则模拟研究进一步探究四个模型的心理计量学性能。模拟研究 1 主要探究四个模型在不同模拟条件下的参数估计返真性；模拟研究 2 主要探究四个模型对数据缺失比例的耐受性。

4.1. 研究 1：参数返真性

4.1.1. 研究设计与数据生成

研究 1 中测验时间点数量固定为 $P = 5$，另外包含 4 个操纵变量，分别是样本量 $N = 100$ 和 300，每个时间点测验长度 $I_p = 15$ 和 30，相邻时间点潜在加工速度的均值增幅$\Delta\mu = 0.25$ 和 0.5，以及各时间点潜在加工速度的方差$\boldsymbol{\sigma}_\tau^2$ = 无变化$(1, 1, 1, 1, 1)^{\mathrm{T}}$、线性变化$(1, 1.25, 1.5, 1.75, 2)^{\mathrm{T}}$和非线性变化$(1, 1.1, 1.3, 1.6, 2)^{\mathrm{T}}$。

如图 6 所示，采用锚题设计，设定时间点 p 的后 20%题目(i.e., $I_p = 15$ 时 3 题，$I_p = 30$ 时 6 题)和时间点 $p + 1$ 的前 20%题目为相同锚题(i.e., 共 4 组锚题)。参考相关研究(Fox & Marianti, 2016; 詹沛达等, 2020)，各题目参数按如下分布生成：$\xi_{ip} \sim N(\mu_\xi, \sigma_\xi^2) = N(4, 0.25)$，$\phi_{ip} \sim N(\mu_\phi, \sigma_\phi^2) = N(1, 0.05)$ 和 $\omega_{ip} \sim N(\mu_\omega, \sigma_\omega^2) = N(2, 0.05)$。5 个时间点的题目参数生成后，对于相同锚题而言，再将时间点 $p + 1$ 上锚题的题目参数固定为时间点 p 上锚题的题目参数，如$\xi_{ip} \Rightarrow \xi_{i(p+1)}$。潜在加工速度依多元正态分布生成，各时间点上潜在加工速度的均值和标准差依不同模拟条件而定，各时间点上潜在加工速度之间的相关系数固定为 0.9。

最后，基于各生成数据，依据公式 3 生成各时间点上的观测 RT。每个模拟条件均生成 50 组数据。

$p = 1$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$p = 2$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$p = 3$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$p = 4$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$p = 5$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

图 6 模拟研究 1 锚题设定示意图($I_p = 15$).

注：相同颜色方框表示相同组锚题.

4.1.2. 分析

在不同模拟条件下分别使用模型 MVN-LRTM、MVN-LRTM-M、LGC-LRTM-L 和 LGC-LRTM-N 作为数据分析模型；数据分析过程与实证数据部分保持一致(e.g., 两条马尔可夫链，每条链含 10000 次

迭代，其中预热 5000 次)。采用均方根误差 RMSE 评估参数估计返真性：$\mathrm{RMSE}(\hat{x}) = \sqrt{\sum_{r=1}^{R}(\hat{x}_r - x_r)^2 / R}$，其中，$x_r$ 和 $\hat{x}_r$ 分别第 r (r = 1, 2, ..., R =50)组数据中某单一参数的真值和估计值。

4.1.3. 结果

表 5 呈现了 MVN-LRTM-M 和 LGC-LRTM-L 两模型中各模型参数的均方根误差(限于篇幅原因，四个模型的完整结果汇总详见 https://docs.qq.com/sheet/DTUJoVEhDUG1LSkpR)。在阐述结果之前需说明的是，同一类模型受操作变量水平变化的影响趋势是一致的，正文仅以 MVN-LRTM-M 和 LGC-LRTM-L 进行阐述；另外，操作变量的水平变化对两类模型的影响存在差异。具体结果如下：第一，当样本量增加时，MVN-LRTM-M 的潜在加工速度均值的 RMSE 减小，而 LGC-LRTM-L 的增大；两模型的潜在加工速度的 RMSE 均减小，三个题目参数的 RMSE 均减小。第二，当题目数量增加时，MVN-LRTM-M 的潜在加工速度均值的 RMSE 增大，而 LGC-LRTM-L 的减小；两模型的潜在加工速度的 RMSE 均增大，题目时间强度参数的 RMSE 略减小，题目时间区分度参数增加。第三，当潜在加工速度的均值增幅Δμ增加时，两模型的潜在加工速度均值的 RMSE 均增大，潜在加工速度的 RMSE 均增大，三个题目参数几乎不受影响。第四，不同方差变化类型对两模型中各参数的影响均较小；但对潜在加工速度而言，似乎方差无变化时的 RMSE 最小。第五，LGC-LRTM-L 对潜在加工速度和潜在加工速度均值的 RMSE 普遍小于 MVN-LRTM-M 的；且前者的时间区分度参数的 RMSE 也普遍小于后者的；第六，随着时间发展(i.e., $p = 1 \rightarrow p = 5$)，MVN-LRTM-M 的潜在加工速度均值的 RMSE 增大，而 LGC-LRTM-L 的减小；两模型的潜在加工速度的 RMES 均增大，两模型的题目时间峰度参数均增大，其余参数几乎不受影响。整体而言，模拟研究 1 结果表明四个模型在多种模拟条件下的参数估计返真性良好，且两个 LGC-LRTM 对潜在加工速度的估计精度略高于两个 MVN-LRTM 的。

表 5. 模拟研究 1 中各模型参数的均方根误差汇总(仅呈现 LGC-LRTM-L 和 MVN-LRTM-M).

模型	*N*	*I*	Δμ	方差变化	μ	τ	ξ	φ	ω
MVN-LRTM-M	100	15	0.25	线性	0.119	0.256	0.073	0.261	0.172
				无	0.112	0.227	0.069	0.270	0.176
				非线性	0.120	0.246	0.071	0.262	0.168
			0.5	线性	0.229	0.317	0.082	0.275	0.173
				无	0.220	0.298	0.089	0.269	0.172
				非线性	0.245	0.323	0.078	0.287	0.172
		30	0.25	线性	0.141	0.290	0.060	0.355	0.166
				无	0.131	0.260	0.060	0.361	0.167
				非线性	0.146	0.293	0.057	0.365	0.168
			0.5	线性	0.282	0.380	0.069	0.373	0.172
				无	0.276	0.362	0.074	0.375	0.170
				非线性	0.285	0.375	0.063	0.368	0.166
	300	15	0.25	线性	0.055	0.152	0.049	0.108	0.131
				无	0.061	0.150	0.059	0.129	0.134
				非线性	0.059	0.156	0.056	0.117	0.135
			0.5	线性	0.106	0.179	0.054	0.114	0.137
				无	0.119	0.187	0.058	0.134	0.140
				非线性	0.125	0.201	0.053	0.140	0.131
		30	0.25	线性	0.081	0.193	0.044	0.195	0.134
				无	0.085	0.178	0.046	0.205	0.135
				非线性	0.086	0.191	0.044	0.195	0.133
			0.5	线性	0.165	0.230	0.046	0.196	0.136
				无	0.162	0.226	0.047	0.198	0.139
				非线性	0.167	0.229	0.048	0.194	0.134
LGC-LRTM-L	100	15	0.25	线性	0.070	0.143	0.076	0.080	0.172
				无	0.040	0.119	0.071	0.060	0.176
				非线性	0.059	0.128	0.071	0.062	0.168
			0.5	线性	0.080	0.152	0.086	0.078	0.173
				无	0.061	0.132	0.097	0.057	0.172
				非线性	0.081	0.149	0.078	0.074	0.172
		30	0.25	线性	0.087	0.192	0.062	0.195	0.166
				无	0.058	0.140	0.060	0.139	0.167
				非线性	0.085	0.182	0.055	0.183	0.168
			0.5	线性	0.172	0.244	0.073	0.198	0.172
				无	0.122	0.181	0.073	0.146	0.170
				非线性	0.151	0.215	0.064	0.171	0.166
	300	15	0.25	线性	0.038	0.126	0.048	0.065	0.131
				无	0.029	0.116	0.058	0.050	0.134
				非线性	0.038	0.121	0.056	0.049	0.135
			0.5	线性	0.062	0.135	0.055	0.062	0.136
				无	0.045	0.120	0.062	0.046	0.141
				非线性	0.051	0.127	0.052	0.053	0.131
		30	0.25	线性	0.060	0.154	0.047	0.139	0.134
				无	0.045	0.118	0.044	0.105	0.135
				非线性	0.058	0.138	0.047	0.115	0.133
			0.5	线性	0.124	0.181	0.045	0.139	0.136
				无	0.091	0.145	0.055	0.097	0.139
				非线性	0.098	0.152	0.046	0.108	0.134

注: 所有数值均为 5 个时间点上的均值.

4.2. 研究 2：缺失值影响

4.2.1. 研究设计、数据生成与分析

实证数据中被试在每个时间点上的 RT 有较大比例(80%)的缺失，尽管这是设计缺失，但我们仍想了解新模型对缺失数据的耐受性。因此，研究 2 拟探讨不同数据缺失比例对四个模型的参数估计精度的影响。本研究聚焦于 1 个操作变量，即被试在每个时间点上 RT 的缺失值比例 MS = 0%、20%、40%、60%和 80%。为使研究 2 的结果更具实践意义，其他变量参考实证研究设定：$P = 5$、$N = 350$、$I_p = 50$ (对应 MS 各水平，分别缺失 0、10、20、30 和 40 题)、$\Delta\mu = 0.5$ 和$\boldsymbol{\sigma}_\tau^2 = (1, 1, 1, 1, 1)$。采用与实证研究一致的重复测量设计，50 道题目的参数和 350 名被试的潜在加工速度完全按照实证数据中的估计值设定。在不考虑参数估计惩罚时，LGC-LRTM-N 对数据的拟合相对最好(i.e., -2LL 最小)，因此，我们将 LGC-LRTM-N 在实证研究中得到的题目参数估计值和被试参数估计值视为模拟研究 2 中的相应参数的真值。依据公式 3 生成各时间点上的观测 RT，每个模拟条件均生成 50 组数据。

图 7 显示了每种缺失值比例条件下，50 组生成中所有被试在每道题目(共 250 题)上的平均对数 RT 和实证数据中每道题目上的平均对数 RT 之间的 Lowess 平滑拟合曲线(局部加权多项式回归曲线)(Cleveland, 1981)。每种条件下的 50 条 Lowess 平滑拟合曲线均趋近于对角线(i.e., 两组数据之间具有高线性相关)，表明每种条件下的 50 组生成数据均能很好地代表实证数据(i.e., 对数据缺失的操纵没有对数据其他特征[e.g., 群体均值和方差]产生影响)。

分别使用模型 MVN-LRTM、MVN-LRTM-M、LGC-LRTM-L 和 LGC-LRTM-N 分析数据。数据分析过程及参数估计返真性指标与实证研究和模拟研究 1 中保持一致。

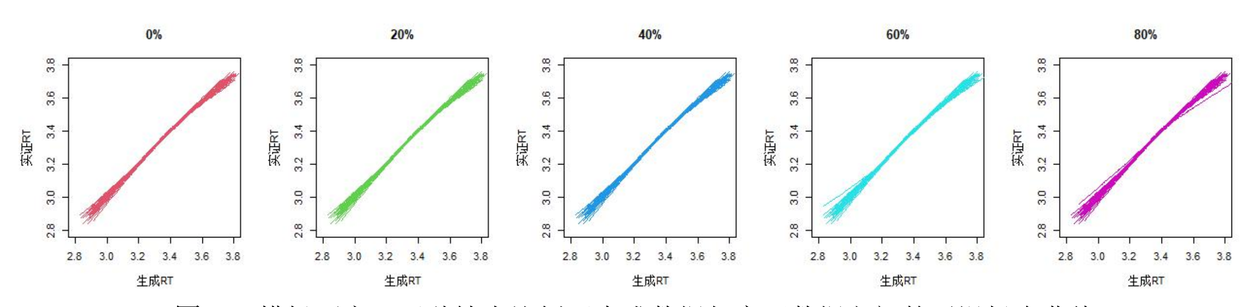


图 7. 模拟研究 2 五种缺失比例下生成数据与实证数据之间的平滑拟合曲线.

4.2.2. 结果

图 8 呈现了研究 2 中各时间点潜在加工速度的和题目参数的估计返真结果。首先，随缺失比例增加四个模型的估计返真性均呈现下降趋势(i.e., RMSE 增大)。其次，缺失比例增加对被试参数返真性的影响大于对题目参数返真性的影响。再有，当缺失比例由 60% → 80%时，各返真性指标会出现一个较大幅度的变化；因此，推荐在实践应用中将完全随机缺失比例控制在 60%以下。最后，结合模拟研究 2 结果，回顾实证研究结果，需要意识到实证研究结果中对被试潜在加工速度的发展轨迹描述可能存在一定的偏差。

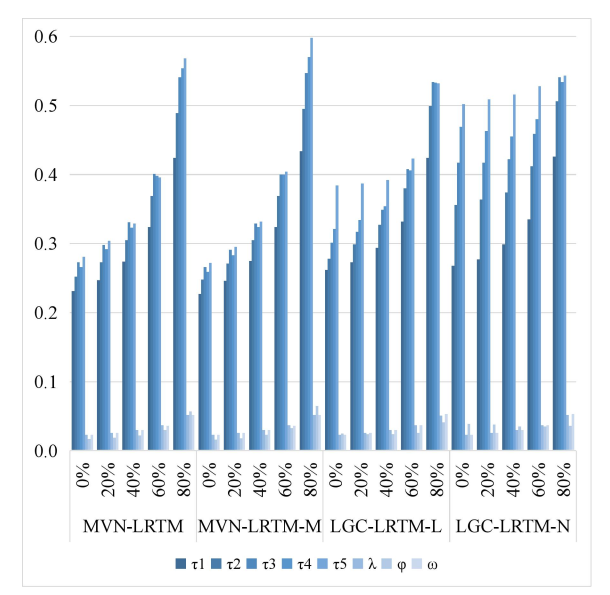


图 8. 模拟研究 2 被试参数的和题目参数的均方根误差.

5. 总结与讨论

为实现对个体潜在加工速度发展的客观追踪，本文基于多元正态分布和潜在增长曲线提出了四个纵向 RT 模型，分别为 MVN-LRTM、MVN-LRTM-M、LGC-LRTM-L 和 LGC-LRTM-N。四个模型的测量模型一致，差异主要体现在描述潜在加工速度如何随时间变化的结构模型上。具体而言，前两个模型直接估计各时间点上的潜在加工速度，未直接关注变化的过程；相反，后两个模型直接估计潜在加工速度随时间的变化(增长)系数，没有直接估计各时间点上的潜在加工速度。实证研究结果表明四个模型均有实践可应用性，且它们的数据分析结果具有较高的一致性。模拟研究 1 表明四个模型在不同模拟条件下的参数估计返真性良好，且两个 LGC-LRTM 对潜在加工速度的估计精度略高于两个 MVN-LRTM 的。模拟研究 2 结果表明四个模型对不同比例的随机 RT 缺失均具有一定的耐受性，建议在实践应用中将完全随机缺失比例控制在 60%以下。总之，本文提出的四个纵向 RT 模型具有实践可应用性，且心理计量学性能良好，不仅丰富了心理与教育测量中纵向 RT 数据的分析方法，也拓展了纵向潜变量模型的应用范围。

但限于精力和能力，本文也有一些局限性有待未来研究做进一步探讨。第一，尽管本文一次性提出了四个纵向 RT 模型，但鉴于纵向数据分析的快速发展，目前还有诸如增长混合建模和多水平增长建模等多种纵向建模方法。未来可尝试在纵向 RT 数据分析中引入更多的纵向建模方法，以期进一步丰富纵向 RT 数据的分析方法。

第二，模拟研究涉及的条件有限，未来可尝试探究更多操纵变量(e.g., 更多更密集的时间点、不同锚题设计)或已涉及变量的更多水平(e.g., 更大样本量)对新模型表现的影响，以期为新模型的实践应用提供更多的理论指导。

第三，本文仅关注单维潜在加工速度随时间的变化，随着测评情境复杂性日益增加，如何追踪多维潜在加工速度(詹沛达等, 2020)随时间的变化也值得关注和探究。

第四，模拟研究 2 仅关注了完全随机缺失 RT 数据对参数估计返真性的影响，并没有进一步探讨其他类型缺失数据(e.g., 非随机缺失)的影响，也没有关注不同缺失数据插补法的表现(e.g., 陈楠, 刘红云, 2015)；同时，模拟研究 2 中不存在由样本流失导致的缺失数据。而纵向研究中样本流失是一种常见现象，未来可探讨该类型缺失数据对新模型参数估计的影响。

第五，为增加模型的普适性，本文没有考虑协变量对潜在加工速度发展的影响。如有需求，未来也可考虑在四个纵向 RT 模型中引入协变量参数，以探究不同协变量对被试潜在加工速度发展的影响。

第六，本研究采用了贝叶斯 MCMC 算法。在贝叶斯参数估计值中，先验分布的选择反映了数据分析者对模型参数的已有经验或信念。根据已有数据分析经验以及已有研究结果(詹沛达等, 2020; Fox & Marianti, 2016; Wang, S., Zhang et al., 2018)，本文选取了特定的先验分布。尽管稳健性分析结果表明新模型的模型参数估计受不同先验分布的影响不大，但这并不意味着本文所用的先验分布适用于所有测验情境。在后续实践中，数据分析者可尝试使用其他先验分布或超先验分布来探索恰当的先验分布。另外，实践者也可尝试使用诸如 M*plus* 等其他软件实现参数估计。

参考文献

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika, 50*, 3–16.

Beck, L. W. (1943). The principle of parsimony in empirical science. *The Journal of Philosophy, 40*, 617–633.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation perspective*. Hoboken, NJ: Wiley-Interscience.

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*(4), 434–455

Cai, L. (2010). High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika, 75*(1), 33-57.

Cleveland, W. S. (1981) LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician, 35*, 54. doi: 10.2307/2683591.

Chang, H.-H., & Wang, W. (2016). "Internet Plus" measurement and evaluation: A new way for adaptive learning. *Journal of Jiangxi Normal University (Natural Science), 40*(5), 441-455.

[张华华, 汪文义. (2016). “互联网+”测评：自适应学习之路. *江西师范大学学报(自然科学版), 40*(5), 441-455.]

Chen, N., & Liu, H. (2015). Comparison of methods addressing MNAR missing data when fitting a latent growth model: Selection model and ML. *Journal of Psychological Science, 38*(2), 446-451.

[陈楠, 刘红云. (2015). 基于增长模型的非随机缺失数据处理: 选择模型和极大似然法. *心理科学, 38*(2), 446-451.]

Chen, Y., Culpepper, S. A., Wang, S., & Douglas, J. (2018). A hidden Markov model for learning trajectories in cognitive diagnosis with application to spatial rotation skills. *Applied Psychological Measurement, 42*, 5 - 23.

Collins, L. M., & Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*. New York: Wiley.

Curtis, S. M. (2010). BUGS code for item response theory. *Journal of Statistical Software, 36*(1), 1-34.

de Boeck, P., & Jeon, M. (2019). An overview of models for response times and processes in cognitive tests. *Frontiers in Psychology, 10*, 102

de Haan-Rietdijk, S., Kuppens, P., Bergeman, C. S., Sheeber, L. B., Allen, N. B., & Hamaker, E. L. (2017). On the use of mixed Markov models for intensive longitudinal data. *Multivariate behavioral research, 52*(6), 747-767.

Embretson, S. E. (1991). Implications of a multidimensional latent trait model for measuring change. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 184–197). Washington, DC: American Psychological Association.

Ferrando, P. J., & Lorenzo-Seva, U. (2007). A measurement model for Likert responses that incorporates response time. *Multivariate Behavioral Research, 42*(4), 675–706.

Fox, J.-P., & Marianti, S. (2016). Joint modeling of ability and differential speed using responses and response times. *Multivariate Behavioral Research, 51*(4), 540-553.

Guo, J., Xu, X., Ying, Z., & Zhang, S. (2021). Modeling not-reached items in timed tests: A response time censoring approach. *Psychometrika*, 1-33.

Guo, L. Shang, P., & Xia, L. (2017). Advantages and illustrations of application of response time model in psychological and educational testing. *Advances in Psychological Science, 25*(4), 701–712.

[郭磊，尚鹏丽，夏凌翔. (2017). 心理与教育测验中反应时模型应用的优势与举例. *心理科学进展, 25*(4), 701–712.]

Han, Y., Xiao, Y., Liu, H. (2022). Feature extraction and ability estimation of process data in the problem-solving test. *Advances in Psychological Science, 30*(6), 1393-1409.

[韩雨婷，肖悦，刘红云. (2022). 问题解决测验中过程数据的特征抽取与能力评估. *心理科学进展, 30*(6), 1393-1409.]

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Newbury Park, CA: Sage Publications.

Klein Entink, R. H., Fox, J.-P., & van der Linden, W. J. (2009). A multivariate multilevel approach to the modeling of accuracy and speed of test takers. *Psychometrika, 74*(1), 21–48.

Klein Entink, R. H., van der Linden, W. J., & Fox, J.-P., (2009). A box-cox normal model for response times. *British Journal of Mathematical and Statistical Psychology, 62*, 621–640.

Kohli, N., & Harring, J. R. (2013). Modeling growth in latent variables using a piecewise function. *Multivariate Behavioral Research, 48*(3), 370-397.

Levy, R., & Mislevy, R. J. (2016). *Bayesian psychometric modeling*. Boca Raton, FL: CRC Press.

Liu, H., & Meng, Q. (2003). A review on longitudinal data analysis method and it's development. Advances in Psychological Science, *11*(5), 586-592

[刘红云，孟庆茂. (2003). 纵向数据分析方法. *心理科学进展, 11*(5), 586-592.]

Liu, Y., Xu, H., Chen, Q., & Zhan, P. (2022). The measurement of problem-solving competence using process data. *Advances in Psychological Science, 30*(3), 522-535.

[刘耀辉，徐慧颖，陈琦鹏，詹沛达. (2022). 基于过程数据的问题解决能力测量及数据分析方法. 心理科学进展, 30(3), 522-535.]

Liu, Y., Du, H., Fang, J., & Wen, Z. (2022). Methodology study and model development for analyzing longitudinal data in China's mainland. *Advances in Psychological Science, 30*(6), 1-13.

[刘源，都弘彦，方杰，温忠麟. (2022). 国内追踪数据分析方法研究与模型发展. 心理科学进展, 30(6), 1-13.]

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika, 55*, 107-122.

Muthén, B., & Muthén, L. K. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and experimental research, 24*(6), 882-891.

Paek, I., Li, Z., & Park, H. (2016). Specifying ability growth models using a multidimensional item response model for repeated measures categorical ordinal item response data. *Multivariate Behavioral Research, 51*, 569–581.

Piech, C., Spencer, J., Huang, J., Ganguli S., Sahami, M., et al. (2015). *Deep knowledge tracing*. arXiv: 1506.05908. https://doi.org/10.48550/arXiv.1506.05908

Plummer, M. (2015). Jags: Just another Gibbs sampler (version 4.0.0). Retrieved from http://mcmc-jags.sourceforge.net/

Tang, F., & Zhan, P. (2021). Does diagnostic feedback promote learning? Evidence from a longitudinal

cognitive diagnostic assessment. *AERA Open*, 7.

van der Linden, W. J. (2006). A lognormal model for response times on test items. *Journal of Educational and Behavioral Statistics, 31*(2), 181−204.

von Davier, M., Xu, X., & Carstensen, C. H. (2011). Measuring growth in a longitudinal large-scale assessment with a general latent variable model. *Psychometrika, 76*, 318–336.

Wang, C., Chang, H., & Douglas, J. (2013). The linear transformation model with frailties for the analysis of item response times. *British Journal of Mathematical and Statistical Psychology, 66*, 144–168. https://doi.org/10.1111/j.2044-8317.2012.02045.x

Wang, C., & Nydick, S. W. (2020). On longitudinal item response theory models: A didactic. *Journal of Educational and Behavioral Statistics, 45*(3), 339-368.

Wang, L., Tang, F., & Zhan, P. (2020). Effect analysis of individualized remedial teaching based on cognitive diagnostic assessment: Taking "linear equation with one unknown" as an example. *Journal of Psychological Science, 43*(6), 1490−1497.

[王立君，唐芳，詹沛达. (2020). 基于认知诊断测评的个性化补救教学效果分析：以"一元一次方程"为例. *心理科学, 43*(6), 1490−1497.]

Wang, S., Hu, Y., Wang, Q., Wu, B., Shen, Y., & Carr, M. (2020). The development of a multidimensional diagnostic assessment with learning tools to improve 3-D mental rotation skills. *Frontiers in Psychology, 11*:305.

Wang, S., Yang, Y., Culpepper, S. A., & Douglas, J. A. (2018). Tracking skill acquisition with cognitive diagnosis models: A higher-order, hidden Markov model with covariates. *Journal of Educational and Behavioral Statistics, 43*, 57–87.

Wang, S., Zhang, S., Douglas, J., & Culpepper, S. (2018). Using response times to assess learning progress: A joint model for responses and response times. *Measurement: Interdisciplinary Research and Perspectives, 16*(1), 45–58.

Wang, S., Zhang, S., & Shen, Y. (2019). A joint modeling framework of responses and response times to assess learning outcomes. *Multivariate Behavioral Research,* 55, 49–68.

Wen, Z. (2017). Causal inference and analysis in empirical studies. *Journal of Psychological Science*, *40*(1), 200−208.

[温忠麟. (2017). 实证研究中的因果推理与分析. *心理科学, 40*(1), 200−208.]

Woolf, B. P. (2009). *Building intelligent tutoring systems*. Morgan Kaufman, Burlington

Zhan, P. (2020). A Markov estimation strategy for longitudinal learning diagnosis: Providing timely diagnostic feedback. *Educational and Psychological Measurement, 80*(6), 1145−1167.

Zhan, P. (2018). *Bayesian cognitive diagnosis modeling incorporating time information: joint analysis of response times and response accuracy data* (Unpublished doctoral dissertation). Beijing Normal University.

[詹沛达. (2018). *引入时间信息的贝叶斯认知诊断建模：对作答时间和作答精度数据的联合分析*(博士学位论文). 北京师范大学.]

Zhan, P. (2022). Joint-cross-loading multimodal cognitive diagnostic modeling incorporating visual fixation counts. *Acta Psychologica Sinica, 54*(4), 1-23.

[詹沛达. (2022). 引入眼动注视点的联合-交叉负载多模态认知诊断建模. *心理学报, 54*(4), 1-23]

Zhan, P., Jiao, H., Man, K. (2020). The multidimensional log-normal response time model: An exploration of the multidimensionality of latent processing speed. *Acta Psychologica Sinica, 52*(9), 1132-1142.

[詹沛达, Jiao, H., & Man, K. (2020). 多维对数正态作答时间模型：对潜在加工速度多维性的探究. *心理学报, 52*(9), 1132-1142.]

Zhan, P., Jiao, H., Liao, D., & Li, F. (2019). A longitudinal higher-order diagnostic classification model. *Journal of Educational and Behavioral Statistics, 44*(3), 241-281.

Zhan, P., Jiao, H., Man, K, & Wang, L. (2019). Using JAGS for Bayesian cognitive diagnosis modeling: A tutorial. *Journal of Educational and Behavioral Statistics, 44*(4), 473–503

Zhang, S., & Chang, H.-H. (2016). From smart testing to smart learning: How testing technology can assist the new generation of education. *International Journal of Smart Technology and Learning, 1*, 67-92.

中共中央，国务院. (2020). *深化新时代教育评价改革总体方案*. URL: http://www.gov.cn/zhengce/2020-10/13/content_5551032.htm

Longitudinal Item Response Times Models for Tracking Change in Latent Processing Speed

Abstract

In psychological, educational, and behavioral studies, measuring change over time is essential to developmental study. These changes can sometimes be captured by longitudinal latent variable models, such as longitudinal item response theory models and latent growth curve models. With the spread of computerized (or web-based) assessments, it has become common to collect process data such as item response time (RT) in addition to traditional item response accuracy (RA) data. RT data is used as a complement to RA data, describes the total time taken by individuals to solve problems and can be used to analyze the latent processing speed of individuals. However, a review of the existing studies reveals that existing longitudinal models focus on longitudinal RA data and lack attention to longitudinal RT data; Moreover, most of the existing RT models are limited to analyzing cross-sectional RT data and cannot track the development of students' latent processing speed over time. To this end, four longitudinal RT models based on two commonly used longitudinal modeling methods (i.e., multivariate normal distribution modeling and latent growth curve modeling) were proposed to achieve objective tracking of individual potential processing speed development and enrich the analysis methods of longitudinal RT data.

Based on the most commonly used cross-sectional RT model, the lognormal RT model (LRTM), four longitudinal RT models were proposed, including the multivariate normal distribution-based LRTM (denoted as MVN-LRTM) and its constraint model with the Markov property (denoted as MVN-LRTM-M), the linear latent growth curve-based LRTM (denoted as LGC-LRTM-L), and the nonlinear latent growth curve-based LRTM (denoted as LGC-LRTM-N). The measurement models are consistent across the four models, with differences mainly in the structural model describing how the latent processing speed changes over time. First, an adaptive learning/assessment dataset about spatial rotation ability was used as an empirical example to show the practical applicability of the proposed models. Second, two simulation studies were conducted further to explore the psychometric performance of the proposed models. The purpose of simulation study 1 was to explore the recovery of parameter estimation under different simulated conditions. The purpose of simulation study 2 was to explore the tolerance of the proposed models to different proportions of missing RT data.

The results of the empirical study mainly indicated that all four longitudinal RT models are practically applicable and have high consistency in the analysis results for the same cohort of data. The results of simulation study 1 showed that the parameters of the proposed models can be well recovered under various simulated conditions. The results of simulation study 2 mainly indicated that the proposed models are tolerant to different proportions of missing RT data, and it was suggested that the proportion of missing RT data should be controlled below 60% in practical applications.

Overall, the four longitudinal RT models proposed in this paper have practical applicability and good psychometric performance, which enriches the analysis of longitudinal RT data in psychological and

educational assessments.

Keywords: response times; longitudinal data; lognormal response times model; latent growth curve; item response theory

附录:

S1. 先验分布对参数估计的影响

S1.1. 低、中和高信息先验分布

S1.1.1. 中信息先验分布

在贝叶斯 MCMC 参数估计中，先验分布的选择反映了数据分析者的经验和对模型参数的预判。根据已有分析经验和研究结果(詹沛达等, 2020; Fox & Marianti, 2016; Wang, S., Zhang et al., 2018)，中信息先验分布(i.e., 包含适量信息的先验分布)设定如下：

首先，所有模型的题目参数的先验分布一样：

$$\xi_i \sim \mathrm{N}(4, 1),\ \phi_i \sim \mathrm{N}(1, 1)\ \mathrm{I}(\phi_i > 0), \omega_i \sim \sqrt{\mathrm{InvGamma}(1, 1)}.$$

其次，两个 MVN-LRTM 而言，潜在加工速度均值的先验分布为：

$$\mu_{p\geq 2} \sim N(0, 1).$$

对 LGC-LRTM-L 而言，增长曲线系数的先验分布为：

$$\mu_{\pi_1} \sim N(0, 1).$$

对 LGC-LRTM-N 而言，增长曲线系数的先验分布为：

$$\mu_{\pi_1} \sim N(0, 1)\text{和}\mu_{\pi_2} \sim N(0, 1).$$

S1.1.2. 低信息先验分布

低信息先验分布以大方差为变异范围，在 S1.1.1 的基础上，低信息先验分布设定如下：

$$\xi_i \sim \mathrm{N}(0, 10),\ \phi_i \sim \mathrm{N}(0, 10)\ \mathrm{I}(\phi_i > 0), \omega_i \sim \sqrt{\mathrm{InvGamma}(10, 10)}.$$

$$\mu_{p\geq 2} \sim N(0, 10).$$

$$\mu_{\pi_1} \sim N(0, 10), \mu_{\pi_2} \sim N(0, 10).$$

S1.1.3. 高信息先验分布

高信息先验分布围绕“真值”设定，并以小方差为变异范围，在 S1.1.1 的基础上，高信息先验分布设定如下：

$$\xi_i \sim \mathrm{N}(4, 0.5),\ \phi_i \sim \mathrm{N}(0.25, 0.5)\ \mathrm{I}(\phi_i > 0), \omega_i \sim \sqrt{\mathrm{InvGamma}(2, 6)}.$$

$$\mu_{p\geq 2} \sim N(0.5(p-1), 0.5).$$

$$\mu_{\pi_1} \sim N(0.5, 0.5), \mu_{\pi_2} \sim N(0, 0.5).$$

S1.2. 参数估计一致性

选用正文所用实证研究数据，该数据包含较大比例缺失值，意味着参数估计结果更易受到先验分布的影响，因此更适合用于探究模型参数对先验分布的敏感性。模型的参数估计设定与实证研究保持一致。图 S1 呈现了四个模型在不同信息量先验分布下的题目参数估计值。图 S2 呈现了四个模型在不同信息量先验分布下的潜在加工速度之间的相关散点图(由于篇幅限制，对 5 个时间点的潜在加工速度求均值)。整体而言，当采用包含不同信息量的先验分布时，每个模型的参数估计结果均无明显变化、较为稳定，即模型对不同信息量先验分布具有较高的稳健性。

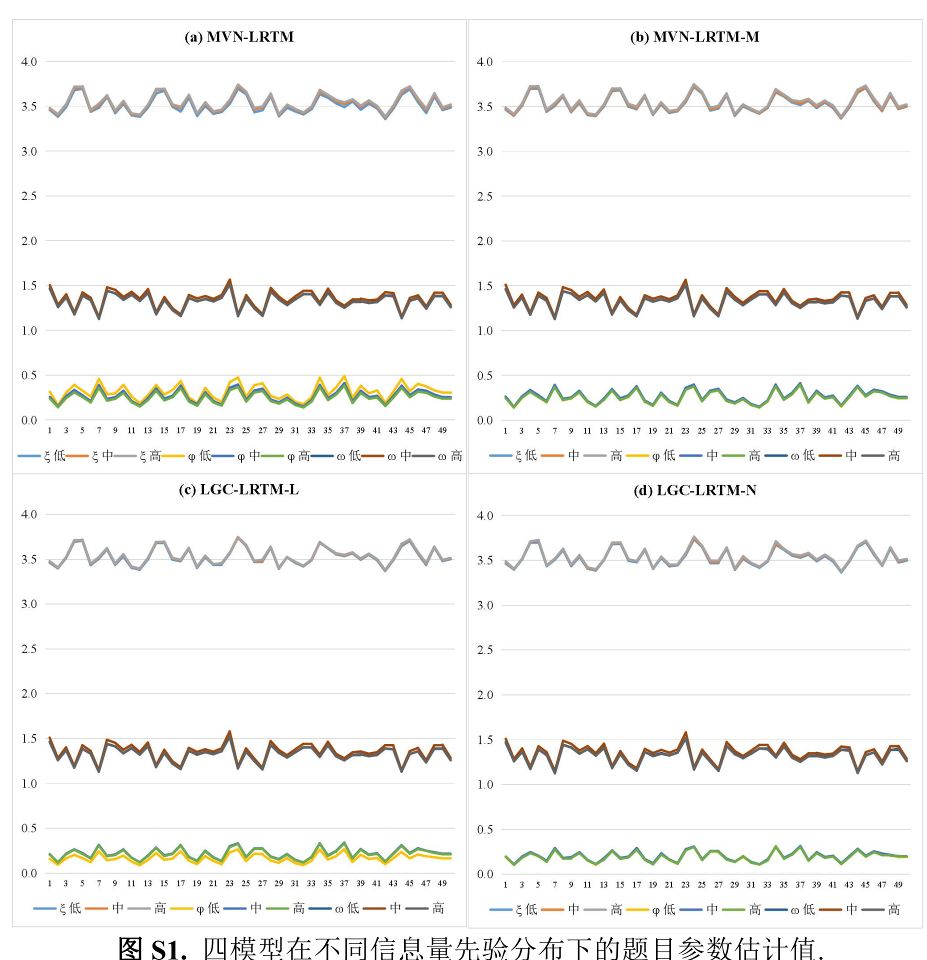


图 S1. 四模型在不同信息量先验分布下的题目参数估计值.

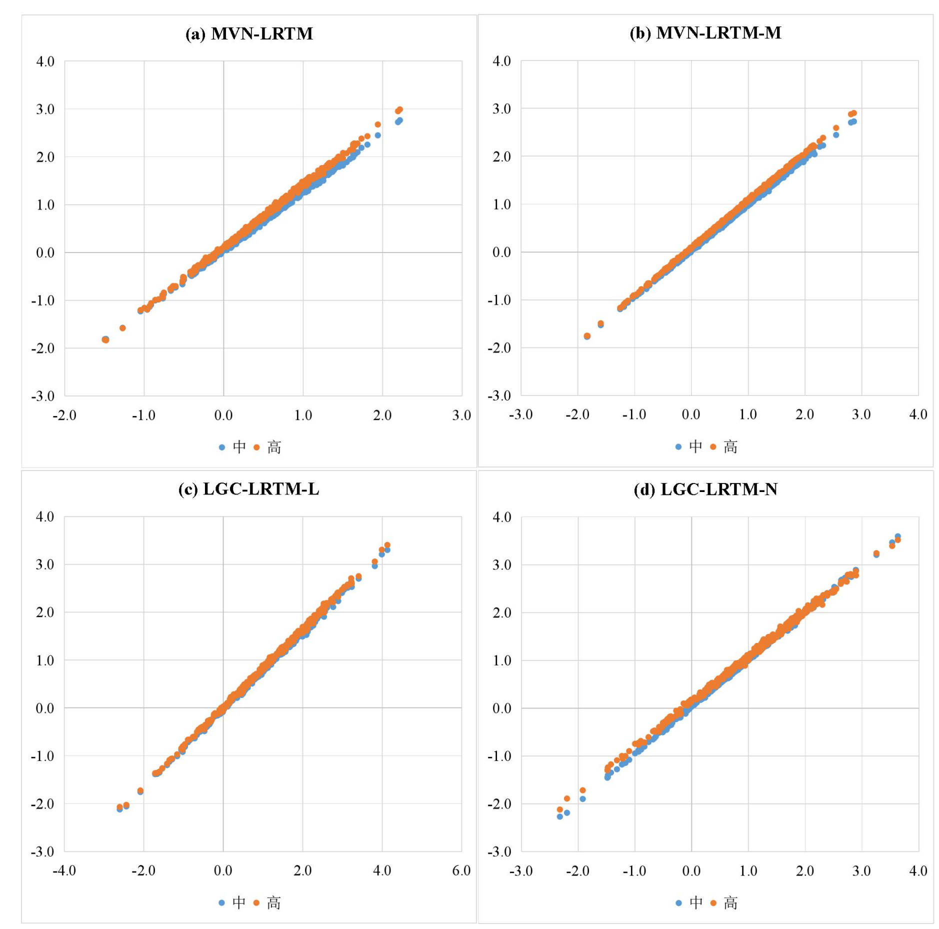


图 S2. 四模型在不同信息量先验分布下的潜在加工速度估计值之间的相关散点图(以低信息量为 x 轴).